Avoiding Overconfidence in Predictions of Residential Energy Demand through Identification of the Persistence Forecast Effect

Huseyin Burak Akyol, Chris Preist, and Daniel Schien

Abstract—**Forecasting domestic electricity consumption is important for a wide range of modern power system solutions and smart applications that support network operation, grid stability, and demand-side management, most of which depend on robust and accurate predictions. The methods producing these predictions infer future load from statistical regularity in historical data. If such regularity is lacking, predictions then regress towards the most recently observed consumption value used in the input set. Predictions then follow the actual load data one step behind in time, potentially affecting the robustness of predictions and functionality of applications. Current evaluation methods do not detect this behaviour which may result in overconfidence in prediction results. In this study, we I) define and systematically analyse this behaviour, which we label the Persistence Forecast Effect and illustrate its impacts, II) propose a novel method, called 1-Step-Shifting, to detect its presence, and III) analyse and establish the relationship between irregularity in data and the effect. Further, we provide a case study applying state-of-the-art forecasting techniques to a real-world dataset of electricity consumption data from 69 households in order to demonstrate the Persistence Forecast Effect, its implications, and its relationship to statistical regularity in historical data.**

Index Terms—**Demand forecasting, Demand-side management, Energy consumption, Energy efficiency, Time-series analysis.**

I. Introduction

With the integration of information and communication technologies and advanced metering infrastructures, new opportunities arise to improve our capacity to understand and efficiently manage electricity consumption at all levels of the electricity grid, including households [1].

In most countries, household electricity consumption accounts for a significant proportion of the overall demand [2], [3]; and many opportunities remain for efficiency improvements and alleviation of the volatile load that the households bring into the power systems [4]. To this end, households are increasingly equipped with smart sensors, storage devices, and distributed and renewable energy generators to optimise consumption and minimise waste through intelligent and automatic energy management systems [5], [6]. Such systems, which enable household energy users to better manage their consumption, and system operators to maintain the grid reliability and supply/demand balance, inevitably rely on accurate and robust consumption forecasts [1], [7]. However, achieving accurate and robust consumption forecasts at the household level is considerably difficult owing to the high volatility and the lack of regularity in load demand [8], [9] that arise from a number of complex factors, such as dweller's lifestyle, income, cultural background, occupancy, location, weather conditions and more [10], [11]. This volatility can adversely affect the robustness and reliability of forecasts [1], [9], [12] as time-series prediction methods infer future electricity consumption from historical load patterns [13], [14]. Note - we use the terms prediction and forecast interchangeably in this text.

In case such regularity is lacking, we observe for state-of-the-art forecasting methods that they produce forecasts that approximate the most recently observed value which results in a time-series of predicted values that is nearly identical to the time-series of observed values, but systematically delayed one step in time, similar to a persistence model [15]. We have labeled this the "Persistence Forecast Effect" (PFE) and explain it in detail in the Problem Statement section. To the best of our knowledge, this phenomenon has so far not been described or analysed in the literature. The impact of the PFE on applications varies between applications depending on their tolerance to accommodate temporal displacement of predictions. The PFE can jeopardise the functionality of such smart applications that require precise timing of predictions since the unnoticed PFE can result in overconfidence in predictions. For instance, battery charging/discharging scheduling and load shifting are important strategies for peak shaving [7], [12], and such strategies require precise forecasts with no temporal flexibility. However, in the presence of the PFE, forecasts are naïvely reproducing the current load, causing batteries to be charged/discharged at a suboptimal time or tasks to be incorrectly scheduled, ultimate resulting in potentially exacerbated peaks, rather than shaving them; a behaviour also observed by [16] for other kinds of delays in predictions of peaks. The PFE is a risk to all prediction contexts that exhibit volatile and uncertain load patterns, including small grids (also called "weak grids") [12]. Hence, it is quite likely to observe the PFE in the predictions for such grids, which can negatively affect energy suppliers' decision-making for strategies with limited temporal flexibility, such as dynamic pricing, tariffs adjustment, supply-demand balancing, and network level peak reduction. However, the state-of-the-art evaluation metrics used in demand forecasting literature are not able to detect

This work was partially supported by the Republic of Turkiye Ministry of National Education.

The authors are with the School of Computer Science, Electrical and Electronic Engineering, and Engineering Maths, University of Bristol, Bristol, U.K. (e-mail: h.burak.akyol@bristol.ac.uk (h.burak.akyol@yandex.com); chris.preist@bristol.ac.uk; daniel.schien@bristol.ac.uk)

the PFE. To avoid overconfidence in prediction results, the PFE should be taken into consideration regardless of the aggregation level or building type, so that steps to mitigate it can be taken; including reviewing the input feature set. Otherwise, it can undermine the robustness of methodological decisions made, and negatively affect the final applications and intelligent energy management systems, which may lead to bigger troubles in smart grid concepts adventitiously.

The main contributions of our work are:

i) Definition and description of the PFE are provided.
ii) Current evaluation metrics are reviewed relative to the PFE.
iii) A novel method for detecting the PFE is proposed.
iv) Underlying causes of the PFE are highlighted.
v) The PFE is investigated for both single-step and multi-step forecasts.

The remainder of the paper is structured as follows. In the next section, we define and explain the PFE, followed by a review of related literature. We then propose a method for detecting the PFE and introduce the experimental setup. After that, we present experimental results that illustrate the PFE and its implications. This is followed by a discussion of how the PFE manifests itself in multi-step forecasts. Then, we evaluate the association between the irregular load pattern and the PFE along with the importance of having historical energy data in input set. Finally, we present our conclusions.

II. Problem Statement

In this section, we describe the PFE in single-step forecasts and motivate its identification and evaluation. Further, we explain the implications of the effect.

Electric load forecasting is a time-series forecasting problem. Time-series forecasting can be performed for single-step or multi-step ahead. Single-step forecasts consists of prediction the value of the next time step (y_{t+1}) only, while multi-step forecasts is the task of predicting a range of sequential future values (y_{t+i}, $i = 1, 2, \ldots, H$, where H is the absolute forecasting horizon). At present, in order to solve this problem, various prediction methods exploiting correlations and similarities [13] are applied. These methods utilise a variety of different "features" on which the values in the output domain are assumed to depend (contingent). Most of the time, these features include a certain number of historical data (past observations of electricity consumption), for example in [9], [11], [14]. More formally, we can describe historical electric data as follows; given a time t, E_t denotes the most recent observed electricity consumption value and E_{t+1} refers to the electricity consumption value to be predicted. Hence, the historical load data can be expressed by E_{t-i}, $i = 0, 1, \ldots, K$, where K is the number of data points.

Even though historical data from the output domain can be a very powerful predictor of future values, they may lead to the PFE. In this study, we describe a phenomenon of PFE that can be observed when regularity in historical load data among input features is sufficiently weak. In such a case, predictions (E_{t+1}) approximate the most recently observed electricity demand value (E_t). An example from the dataset in our case study is given in Fig. 1, showing observed values in blue and time-series predictions suffering from the PFE in orange. The main characteristic is that the shape of the predictions is almost identical to the observed values, except that the output domain values are displaced by one time step to the right (the future). In other words, the method returns almost the same value with *observed* value as its *prediction* ($E_t \approx E_{t+1}$). Our succinct explanation for this in due to the volatility and pattern irregularity in historical data, methods cannot learn enough and instead, extrapolate from the most recent electricity demand value because of the high correlation between the consecutive data points in electricity load data.

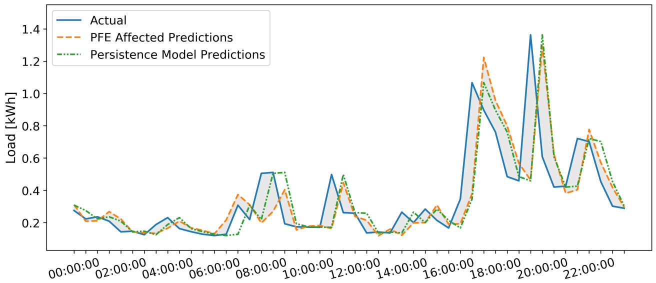


Fig. 1. A comparison of PFE-affected predictions produced by a machine learning method and predictions produced by a a naïve persistence model throughout a day. The error of the predictions affected by the PFE is shaded in grey. The naïve persistence model performs better than the machine learning method: The absolute error of the machine learning method is 6.723 kWh, while the error of the persistence model is lower at 6.526 kWh.

Also shown in Fig. 1 is the output of a persistence model. This is a well-known naïve method that is mostly used as a baseline method for testing the prediction ability of machine learning algorithms [15], [17]. It simply returns the value of the most recent observation (E_t) as a prediction outcome for the next time step (E_{t+1}), resulting in ($E_{t+1} = E_t$). Based on the similarity between the predictions of the persistence model and the predictions suffering from the bias we investigate in this text, we have chosen the term "Persistence Forecast Effect" to describe the effect. In this randomly chosen example, the naïve persistence model has a lower absolute error than the LSTM RNN model.

Evaluation metrics assess the prediction error from a discrepancy between the predicted and observed values. However, the PFE might result in overconfidence in evaluation metric results because the most popular evaluation metrics, e.g. Mean Squared Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE) are not capable to identify this effect by themselves. Smart systems or applications built on top of forecasts affected by the PFE will not perform as effectively as they would with data that enables the developments of models that surpass the naïve persistence model. According to [18], a 1% rise in the prediction error translates to a roughly £10 million increase in annual operating costs for the United Kingdom in 1984. Given the increased level of automation in all parts of the economy, building these smart and automated systems on top of predictions with an unnoticed PFE will likely cost much higher today.

Evaluation metrics can also be used to provide an *in-between* comparison of alternative prediction methods. However, besides being a sign of poor time-series prediction generally, the PFE is also a direct threat to the validity of comparative studies that rank methods in-between. As we will demonstrate in the Results section, the PFE can result in inconsistent rankings or even rank reversal between prediction methods. This in turn threatens the transferability of findings from studies aiming to optimise forecasting methods or compare forecasting methods for a certain type of problem.

III. Related Works

Time-series forecasts are defined by both timings as well as amplitudes of series of events. Therefore, it is critically significant to predict the timing of an event correctly along with its amplitude. An error resulting from an event predicted to happen too early or too late is called phase error [19], [20]. In [21], [22], authors effectively illustrated that the phase error, together with bias and amplitude error, is one of the three main components forming RMSE and MSE. As a consequence, in order to achieve relatively better metric value, the phase error should also be handled properly. In one line of work, studies have focused on forecasting ramp events, referring to sudden, significant fluctuations in time-series data within a short period of time [20]. These aim to predict the time of ramp events correctly alongside the accurate amplitude. More recently, a growing interest has emerged in this topic in solar (see [23]–[25]) and wind (see [26]–[28]) power production domains to guard from the negative impacts of ramp events for greater safety, stability, and economics of power systems and energy storage devices. Further, [17] proposes a novel evaluation metric, Ramp Score, to evaluate the ability to predict significant ramp events.

Furthermore, considering the phase error, the point-wise metrics, such as MAPE, RMSE, and MSE, are claimed to be inappropriate in time-series prediction evaluation [5], [16], [17], [29]. Point-wise metrics simply compare the observed and predicted values at each time step, and hence, they lead to double penalty effect (DPE). The DPE refers to the case in which point-wise metrics penalise temporally displaced predictions twice: first, where the event actually should be, and second, where the event is predicted to happen - even if the size and the amplitude of an event are correctly predicted in principle. In order to avoid this effect, the general recommendation is to tolerate and not penalise small and discontinuous displacement of predictions in time. The authors of [16] introduce an adjusted p-norm error measure that allows for small and discontinuous displacement of predictions in time. Basically, the main idea behind this method is to partly drop the time dimension and provide temporal flexibility to predictions during evaluation. Based on the test they run, they find that the new metric they proposed is suitable and useful for volatile and irregular data while the standard point-wise metrics are adequate for smooth and regular data. Similarly, alignment-based metrics such as Dynamic Time Warping, Longest Common Sequence, Parameterised Forecast Error Metric, and Move Split Merge are also proposed as evaluation metrics for time-series predictions [5], [16]. Alignment-based metrics mainly align the predictions with the actual values in order to find the optimal match between them, hence they do prevent the DPE. However, on the other hand, they do not have consideration of time dimension. That is to say, they do not preserve the time order among data points and each data point is handled as an independent prediction. As a result, they are not suitable for the assessment of electricity consumption prediction and other applications where the time dimension and time order of the data points are inflexible [16].

TABLE I
Recent studies of household electricity forecasts and applied evaluation metrics.

Study	Evaluation Metric	Study	Evaluation Metric
[2]	Corr., MAPE, RMSE	[40]	MAPE, MSE, R-MAPE
		[41]	MAE, RMSE
[4]	RMSE	[42], [43]	MAE
[6], [35], [36]	MAPE, RMSE	[44]	Coef. of Variation
[8]–[10]	MAPE	[45]	MAE, MAPE, MSE
[15]	S-MAPE	[46]	Corr, RMSE
[11], [37], [38]	MAE, MAPE, RMSE	[47]	MAE, Corr., RMSE, MAPE, MaxAE, and SI
[39]	MAE, RMSE, N-RMSE	[48]	Corr., RMSE Coef. of Determination

Due to the uncertainty and volatility in household electricity load and resulting challenges for point forecasts, probabilistic forecasts are considered as a way of getting robust and reliable predictions [29]. A comprehensive review of probabilistic electric load forecasting for various types of buildings and aggregation levels is provided in [30], [31]. Besides, more recently, some studies including [32]–[34] provide new results for probabilistic load forecasting specifically in the residential domain.

Nevertheless, point forecasting methods and point-wise metrics are still predominantly used in household level load forecasting. A number of these works are reviewed from two perspectives below. First, we list some recent studies using point forecasting methods, together with their applied evaluation metrics. We then present certain studies whose results display some level of PFE based on our visual investigation.

Table I lists a number of recent studies, each of which has a different strategy and approach to different variants of household demand forecasting problem, as well as the point-wise evaluation metrics they utilise to evaluate their prediction accuracy. From Table I, it is evident that the most popular metrics in household load forecasting are RMSE and MAPE, applied by 16 and 15 studies, respectively. Furthermore, as seen in Table I, it is a common practice to use multiple evaluation metrics. However, it is of note that justification for the choice of evaluation metrics is rarely given.

Regarding the presence of the PFE, in recently published works, we carried out a visual inspection of plots of predictions and actual observations in several peer-reviewed studies and find substantial evidence for the PFE; for example in Fig. 2 in [10], Fig. 8 in [9], Fig. 9 and 10 in [41], Fig. 8 in [42], Fig. 10 in [44], Fig. 8 in [45], Fig. 3 and 4 in [49], and Fig. 4 and 5 in [50]. However, no comment is made on systematic one step delay in predictions in these studies. Furthermore, many studies, including [37], [43], [47], do not provide plots comparing predicted and actual load. Therefore, it is not possible to claim something definite about the existence or absence of the PFE in such studies.

IV. Methodology

In this section, we describe the novel method we propose for detecting the presence of the PFE in single-step forecasts. We then describe an empirical setup used to illustrate and evaluate the PFE in a large-scale real-world dataset.

A. 1-Step-Shifting Method for PFE Detection

As previously mentioned, the PFE occurs mainly as a result of a lack of regular pattern in data as well as a high correlation between consecutive time-series data points. When the effect arises, forecasting results approximate the value of the previous data point, and hence, they follow the actual energy demand one time step back as illustrated in Fig.1. However, visual inspection alone is not always sufficiently objective for practical and repeatable PFE detection. Given that, a computational approach to detect the PFE is required. Consequently, we propose the "1-Step-Shifting" (1-SS) method. The central idea for the 1-SS method is the recalculation of standard evaluation metrics after shifting the predictions "one step back" in time (shift the predictions to the past) or shifting the time-series of actual observations "one step forward" in time (shift the actual data one time step right to the future). Note that in this text, all the visualisations and formulations utilise the former strategy which shifts the predictions one step back in time to the past. The main idea behind the 1-SS method is to show that shifting the predictions one step back in time yields better evaluation metric results, which proves the systematic one step delay in predictions.

The proposed 1-SS method contains four steps as follow:

Step 1: Calculate evaluation metrics for predictions/actual data as usual.

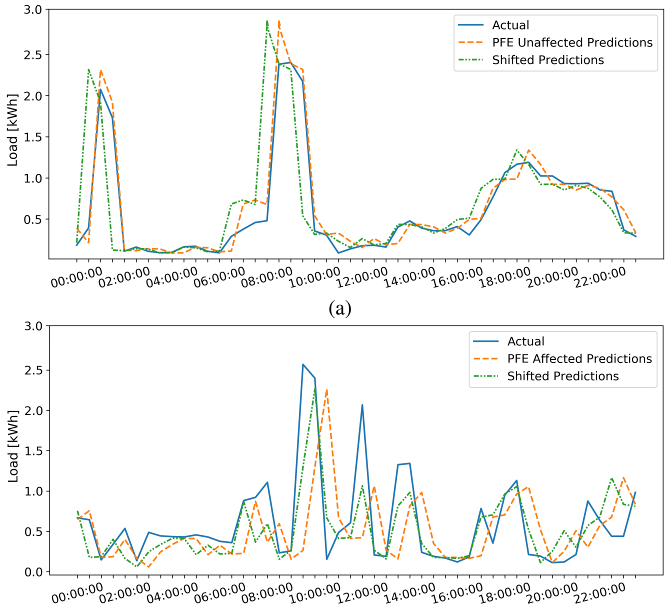
Step 2: Apply shift of predictions one time step to the past (or shift actual load data one step to the future).

Step 3: Recalculate the evaluation metrics for the shifted predictions/actual load data.

Step 4: Compare the evaluation metric results of Step 1 and Step 3. If 1-SS results in considerable improvements in accuracy, then it can be claimed that the predictions exhibit the PFE.

For instance, Fig. 2 illustrates the 1-SS method for predictions affected by the PFE (Fig. 2a) and predictions not affected by the PFE (Fig. 2b). In Fig. 2a, the curve of the predictions matches the curve of the actual values much better than the curve of the shifted predictions do. The opposite is the case in Fig. 2b, where the difference between the shifted predictions to the actual load data is smaller than that of the original predictions.

This four-step 1-SS method is independent of a specific evaluation metric. It is common practice (see Table I) and recommended to apply multiple evaluation metrics [37] as part of a prediction evaluation, as each metric has individual advantages and disadvantages. In this work, therefore, we apply the most popular evaluation metrics in the household load forecasting literature: MAPE, RMSE, and Correlation coefficient. We advocate MAPE and RMSE as they complement each other in many aspects and Correlation as it measures the linear correlation (similarity) between two sets of variables. As a result, these three metrics, whose properties are summarised in Table II, aid us to evaluate the PFE from as many as different angels.



(a)

(b)

Fig. 2. Illustration of the 1-SS method on two predictions; (a) without the PFE and (b) with the PFE.

These three metrics are defined as;

$$MAPE = [1/n \sum_{t=1}^{n} (|Act_t - Pred_t|)/Act_t] \times 100 \quad (1)$$

$$RMSE = \sqrt{1/n \sum_{t=1}^{n} (Act_t - Pred_t)^2} \quad (2)$$

$$Corr = \frac{\left[\sum_{t=1}^{n} (Act_t - \overline{Act})(Pred_t - \overline{Pred})\right]}{\sqrt{\left[\sum_{t=1}^{n} (Act_t - \overline{Act})^2\right]\left[\sum_{t=1}^{n} (Pred_t - \overline{Pred})^2\right]}} \quad (3)$$

where Act_t is the actual value and $Pred_t$ is the forecast value at time t and where $\overline{Act}$ and $\overline{Pred}$ refer to the mean of the actual loads and the mean of the predictions respectively.

On the other hand, 1-SS modifies these formulas as follow;

$$MAPE^* = [1/n \sum_{t=1}^{n} (|Act_t - Pred_{t+1}|)/Act_t] \times 100 \quad (4)$$

$$RMSE^* = \sqrt{1/n \sum_{t=1}^{n} (Act_t - Pred_{t+1})^2} \quad (5)$$

$$Corr^* = \frac{\left[\sum_{t=1}^{n} (Act_t - \overline{Act})(Pred_{t+1} - \overline{Pred})\right]}{\sqrt{\left[\sum_{t=1}^{n} (Act_t - \overline{Act})^2\right]\left[\sum_{t=1}^{n} (Pred_{t+1} - \overline{Pred})^2\right]}} \quad (6)$$

Declaration 1. Given formulas 1-6, we define that for a given household, predictions exhibit the PFE if $MAPE^* < MAPE$, $RMSE^* < RMSE$, and $Corr^* > Corr$. Conversely, predictions do not suffer from the PFE, if $MAPE^* > MAPE$, $RMSE^* > RMSE$, and $Corr^* < Corr$. Any other combinations

TABLE II
PROPERTIES OF MAPE, RMSE AND CORRELATION AS EVALUATION METRICS.

MAPE	RMSE	Correlation
Unit independent (percentage)	Unit dependent (unit of data)	Unit free (-1, 1)
Easy to interpret	Not always easy to interpret	Easy to interpret
It fails if some of the actual values are equal to zero	Not affected if some of the actual values are equal to zero	Measures the strength of the relationship between the relative movements of two sets
Vulnerable to close-to-zero actual values	Vulnerable to extremely high errors	-
Penalises the errors equally (not squares the errors)	More punishment for larger errors (squares the errors)	Reasonable metric to compare the shape and synchronicity of two set of variables
Depends systematically on the level of the time-series	Depends on the magnitude of the error	Depends on the direction of the correlation (whether positive or negative)
Penalises the negative and positive errors equally	Penalises the negative and positive errors equally	-
The smaller MAPE, the better prediction	The smaller RMSE, the better prediction	The higher correlation, the better prediction.

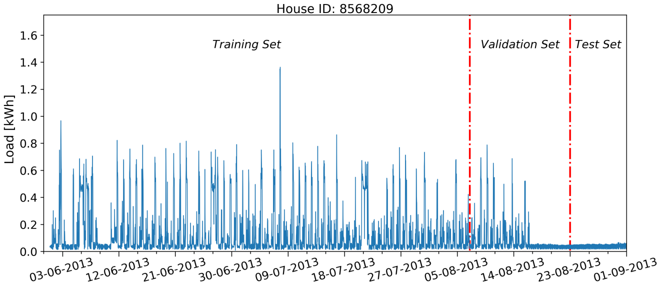


Fig. 3. Load profile of household 8568209 over the train-validation-test split duration.

other than these two, we consider as inconclusive and advise further investigation via manual, visual inspection or in-depth regularity analysis.

B. Experimental Setup

In order to illustrate the PFE and the use of 1-SS for the PFE detection, we carry out an experiment with a large-scale dataset and state-of-the-art machine learning methods. We replicate relevant results from [8] who published prediction results on the publicly available Smart Grid Smart City project dataset (SGSC) [51] and used methods from the Keras library together with the Theano back-end. The dataset they utilise provides electricity consumption data for several houses and they deploy multiple machine learning methods, which are critically important for the purpose of this work. This experiment also allows us to demonstrate that the PFE already exists in literature.

1) Dataset: The SGSC dataset provides energy consumption data for a large number of households in New South Wales, Australia, and captures the variability of residential schedules and activities. We replicate results from [8] and identical to them, utilise a subset of 69 households from the SGSC with 92 consecutive daily load profiles, from a 3-month time span (01.06.2013 - 31.08.2013). This time period was initially chosen as it includes complete half-hourly electric load data for the 69 individual households. We apply the same train-validation-test split of 67 days of data for training, 16 days for validation, and the remaining 9 days for testing.

The input features are identical across all 69 households and include:

- Historical electricity load data for the past 2 time steps (E_t, E_{t-1}).
- Day of week indicator (ranges from 0 to 6).
- Time of day indicator (ranges from 0 to 47).
- Weekend indicator (ranges from 0 to 1).

TABLE III
HYPER-PARAMETER SETTINGS OF THE LSTM RNN AND BPNN METHODS, REPLICATING [8].

Parameter	Setting	Parameter	Setting
Hidden Layers	2	Decay	0.0
Nodes on Each Layer	20	Batch Size	32
Epoch	150	Dropout Rate	No Dropout
Optimizer	Adam	Loss Function	MAE
Learning Rate	0.001		

Identical to [8], we carry out data preparation as well: all the input features are transformed to a standard scale (0, 1) independently, which is crucial to reduce the impact of marginal outliers. To do this, one-hot encoding is applied to time-of-day, day-of-week inputs. Besides that, min-max normalisation is performed for historical electricity data columns.

However, it is of note that for one of the 69 houses (8568209) from the dataset, electricity consumption is approximately zero for a substantial part of the dataset (potentially because the property was vacant), including the entire time span that defines the test set (see Fig. 3). The test set, thus, is utterly dissimilar to the training set. In order to evaluate the PFE for this household, a different train/test split would be needed, which would be a deviation from [8]. For this reason, we have excluded this house from the PFE evaluation.

2) Prediction Methods: In order to illustrate the PFE and its implications, we apply two state-of-the-art machine learning techniques, Long Short-Term Memory Recurrent Neural Network (LSTM RNN) and Back-Propagation Neural Network (BPNN), as applied in [8]. In this section, we introduce their implemented architecture as well as critical hyper-parameters. However, a detailed explanation of their working principles is out of the scope of this study. Interested readers can find an introduction to LSTM RNNs in [4], [8], [39] and to BPNNs in [52], [53].

Both of these methods are built with the same architecture and hyper-parameters, replicating [8]. The common architecture and hyper-parameter settings of the methods are summarised in Table III. The specific values for batch size, dropout rate, and loss function were not stated in [8]. Therefore, we apply the default batch size of 32 and no drop-out. However, the Keras framework does not provide a default choice of loss function, and hence, we carried out a grid search and select the function that resulted in the best reproduction of the results from [8] – which is MAE.

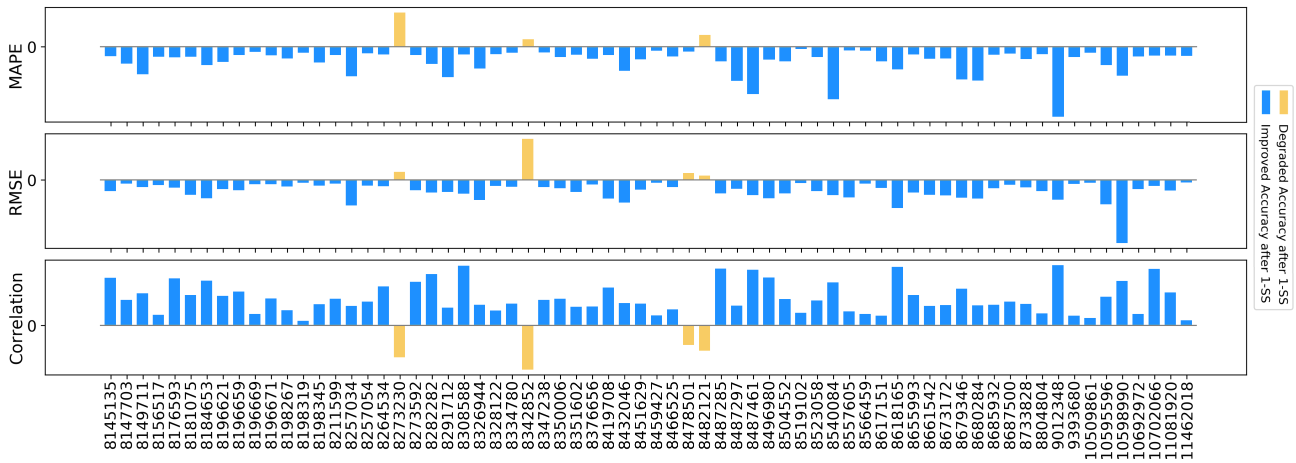


Fig. 4. Difference between default and shifted evaluation metrics with the 1-SS method. Bars are vertically aligned for each household ID. Results indicate that only the predictions of households 8273230, 8342852, 8482121 are PFE-free.

3) Clustering Method: We carry out a clustering analysis of daily load profiles of households in order to explain the association between the presence of PFE and irregularity in load patterns. The clustering provides a visual and numeric comparison of the regularity level of households whose predictions are PFE-free and PFE-affected.

In an electricity demand prediction context, clustering algorithms are generally employed to improve the prediction accuracy [39], [40], [54]. In this text, however, we use clustering for data analysis rather than prediction. The clustering groups the 92 daily load profiles based on their resemblance to one another. The number of clusters provides a measure for the regularity level of data from a household on a day-by-day basis based on hierarchical clustering in line with [8], [9].

Hierarchical clustering (see [55], [56] for details) is chosen, as it I) requires only a few number of hyper-parameter selection, II) does not need a predetermined number of clusters, and III) identifies outliers explicitly.

Clustering is an unsupervised learning method that groups objects based on their relative similarity given some distance metric [54]. In the current context, since the objects are not a single data point but a sequential data array of 48 data points for each day, correlation was chosen as the distance metric. To calculate the distance between individual objects or clusters, the average method is performed. The threshold for assignment to the same cluster is set to $corr = 0.75$. In other words, two daily electricity demand profiles are assigned to the same cluster if their correlation coefficient is equal to or greater than 0.75.

V. Results

We first deploy LSTM RNN on the data of 68 residences independently with identical architecture and hyper-parameters and then, apply the 1-SS method to the predictions produced by the LSTM RNN for 68 residences. The difference between the default evaluation metrics (formulas 1 to 3) and their shifted equivalents (formulas 4 to 6) are shown in Fig. 4.

Considering the difference between default and shifted evaluation metrics as given by the 1-SS method and Declaration 1, the 68 households can be split into three groups:

- PFE-free: All three metrics worsen after 1-SS. This is the case for household IDs: 8273230, 8342852, 8482121.
- Inconclusive: One or two metric(s) improve while another worsens after 1-SS. This is the case only for household ID: 8478501.
- PFE-affected: All three metrics improve after 1-SS. This is the case for the remaining 64 houses, which is by far the majority of households.

For MAPE and RMSE, a larger value corresponds to a worse result, while the opposite is true for Correlation. For instance, the 1-SS method indicates that the predictions of household 8342852 are not affected by the PFE, as values for MAPE and RMSE worsen (error increases) from 38.705 and 0.227 to 46.295 and 0.550, respectively. Meanwhile, the Correlation worsens (less alignment) from 0.941 to 0.643. On the other hand, the predictions of household 8184653 are affected by the PFE, as MAPE and RMSE improve (error decreases) to 20.150 and 0.107 from 38.724 and 0.252, respectively, and the Correlation improves (better alignment) from 0.643 to 0.941.

Even though the 1-SS method is conceptually very simple, it can detect the PFE in all but one of the cases (inconclusive). The default evaluation metrics by themselves are not able to identify such behaviour at all, which can result in misplaced confidence in predictions. Fig. 5 juxtaposes MAPE and MAPE* values of 6 houses in three pairs of houses with PFE-free and PFE-affected predictions. In each pair, the MAPE values are very similar to one another (e.g. 25.357 and 24.815 for houses 8273230 and 8487285, respectively). According to metric results, the prediction method performs relatively similarly for the pairs 8273230 and 8487285, 8342852 and 8184653, and 8482121 and 8661542. However, when the 1-SS method is applied, the evaluation metric values improve for the households with PFE-affected predictions and degrade for those not subject to the PFE – as shown by MAPE* in Fig. 5. Even though the absolute overall prediction error in each pair of households during the test interval is very similar, for predictions exhibiting the PFE, this error is almost exclusively determined by the cumulative difference of the observed energy consumption between two time steps.

Another important risk resulting from PFE is for the interpretation of evaluation metrics during comparison of prediction methods. When using the standard evaluation met-

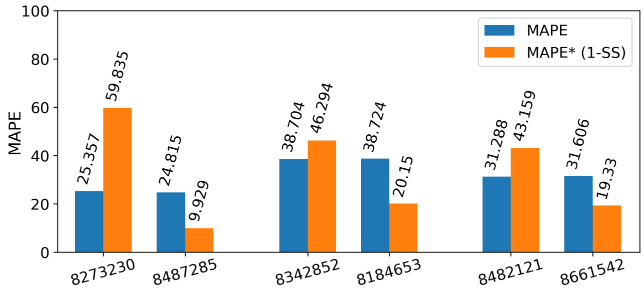


Fig. 5. Paired MAPE and MAPE* values of three PFE-free predictions (8273230, 8342852, 8482121) and three PFE-affected predictions (8487285, 8184653, 8661542).

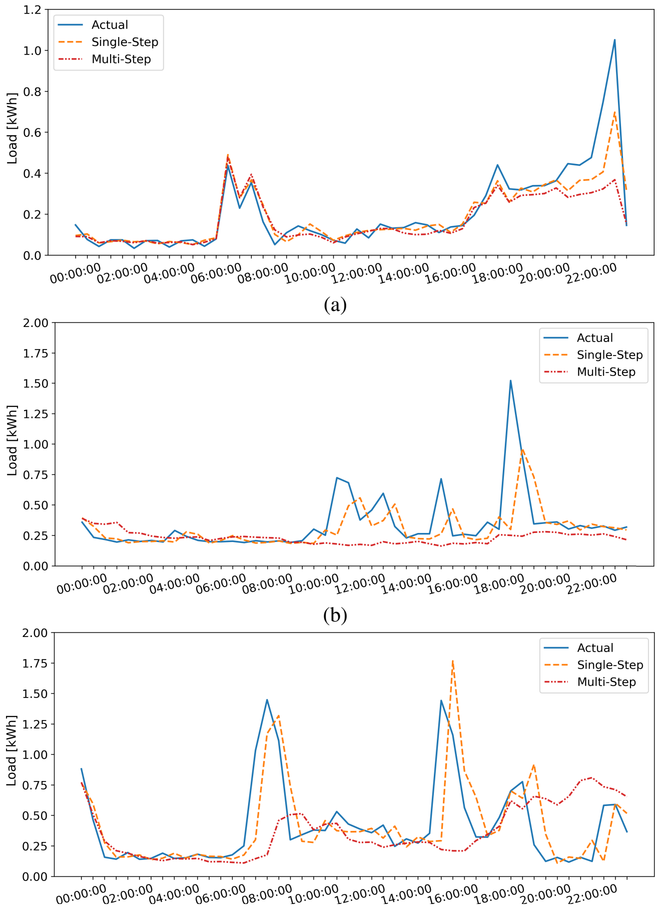
TABLE IV
EVALUATION OF REPEATED RUNS OF LSTM RNN AND BPNN FOR A SELECTION OF HOUSES FROM THE DATASET.

House ID	PFE	Method	MAPE1	MAPE2	MAPE3	MAPE4
8273230	Unaffected	LSTM RNN	25.210	24.617	24.971	25.029
		BPNN	31.084	29.507	32.054	29.876
8342852	Unaffected	LSTM RNN	37.824	38.576	37.560	38.625
		BPNN	63.641	68.311	66.885	65.307
8482121	Unaffected	LSTM RNN	31.303	30.839	30.941	30.471
		BPNN	39.858	40.926	41.360	40.082
8181075	Affected	LSTM RNN	18.920	18.916	18.610	19.104
		BPNN	19.206	18.610	19.778	18.888
8196621	Affected	LSTM RNN	36.807	35.709	36.355	36.765
		BPNN	35.174	36.523	36.624	34.440
11081920	Affected	LSTM RNN	31.785	30.708	31.739	31.671
		BPNN	31.532	30.275	32.569	30.476

rics to compare alternative methods, the PFE can result in evaluation metrics that separate methods only poorly - or even reversing the rank order between them. We illustrate this with a comparison of LSTM RNN and BPNN methods for six residential buildings, three of which have PFE-free predictions and three have PFE-affected predictions. These methods are partly stochastic and result in slightly varying forecasts. We train both methods four times for each of these buildings. The MAPE results are listed in Table IV. For buildings with PFE-free predictions, LSTM RNN performs consistently and significantly better than BPNN. However, for households whose predictions are PFE-affected, the difference between MAPE results is not significant and ranks between LSTM RNN and BPNN are not stable.

A. The PFE in Multi-Step Forecasts

So far, we have only considered single-step forecasts. To investigate the PFE in the context of multi-step forecasts, we deploy LSTM RNN, with the same architecture and hyper-parameters as above, for the same dataset. We perform one-day-ahead prediction, yielding 48 predicted values for the electricity consumption of the following 24 hours. To deal with this task, we use a recursive mechanism that allows us to use our single-step method, LSTM RNN, iteratively. The main idea of the recursive mechanism is to deploy the same pre-trained single-step method for each time point to be predicted. Basically, the single-step model is used to predict the one time point ahead first, and then, the output is fed into the same single-step model to predict the subsequent time point. This procedure is iteratively applied until the last value of the desired multi-step forecasting sequence is predicted.



(a)

(b)

(c)

Fig. 6. Multi-step (day ahead) load predictions of PFE-free House 8482121 (a), and PFE-affected Houses 8487285 (b) and 8661542 (c).

Day ahead multi-step forecasts for three households are presented by Fig. 6. The prediction method produces prediction curves that show no temporal displacement, and the predicted load pattern follows the actual load pattern closely for the building (8482121) whose predictions are PFE-free in a single-step case (Fig.6a). On the other hand, for the buildings (8487285 and 8661542) whose predictions are PFE-affected in single-step case, the method provides forecasts of very poor alignment. The multi-step predictions are either i) more or less flat lines with minimal fluctuations (Fig.6b), or ii) prediction curves are full of arbitrary peaks/troughs (Fig. 6c). In both cases, the forecasts are uncorrelated to the actual load. The correlation coefficients between actual load and multi-step forecasts are -0.123 and 0.026 for forecasts, respectively. Therefore, considering the given definition of the PFE and the dissimilarity between the actual load and multi-step predictions, it does not seem possible to talk about the temporal accuracy or the existence/absence of the PFE for such predictions.

As a result, in any case, the 1-SS method as presented here is not sufficient to detect the PFE in multi-step forecasts, which is an area of further work. However, the 1-SS *can* be used to show the *absence* of the PFE in multi-step forecasts anyway, and can thus be used to give confidence in evaluation outcomes to stakeholders.

TABLE V

HIERARCHICAL CLUSTERING RESULTS FOR 68 RESIDENCES SORTED BY THE NUMBER OF CLUSTERS.

House ID	# of Clus.	House ID	# of Clus.	House ID	# of Clus.	House ID	# of Clus.
8342852	10	8198319	45	8347238	59	8308588	74
8273230	12	8196659	45	8504552	60	8487461	75
8482121	12	8519102	46	8376656	60	8679346	75
8804804	20	8617151	46	8655993	60	8487285	76
11462018	20	8350006	46	8198345	60	8264534	77
8478501	26	8459427	47	8184653	60	8432046	77
8466525	30	10509861	50	8145135	61	8496980	77
8680284	34	8257054	50	8326944	62	10598990	77
8334780	36	8196671	53	8147703	63	8257034	77
8557605	38	8685932	54	8351602	64	9012348	80
10692972	38	8273592	54	8673172	65	11081920	81
8291712	40	8149711	54	8211599	66	10702066	84
8198267	40	9393680	56	8566459	66	8540084	84
8419708	40	8661542	57	10595596	67	8618165	86
8181075	42	8328122	57	8156517	67	8487297	87
8687500	44	8196621	57	8451629	68	8523058	88
8196669	45	8733828	58	8176593	72	8282282	91

B. Clustering Results

Results from hierarchical clustering are shown in Table V. Here, houses are sorted by the number of clusters, with fewer clusters indicating a greater similarity between the daily energy consumption profiles over the 92 days of a specific household. It is important to note that outlier daily demand profiles, that do not have any similar profiles to be in the same cluster with according to the correlation distance metric, are each put in a separate cluster.

The clustering results show significant differences between the regularity level of demand profiles of households. The most regular household has only 10 clusters for the 92 daily load profiles, while, the household that has the most variable load profiles has 91 clusters; in other words only two daily load profiles that were similar.

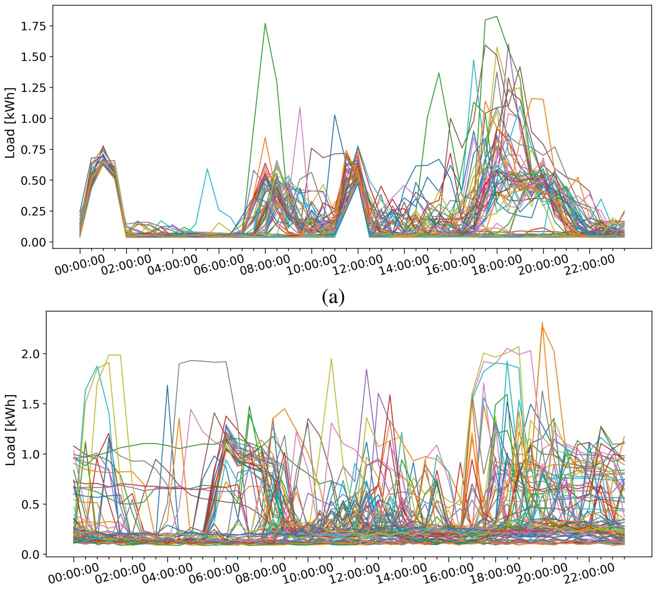
Among the results in Table V, the three houses that have PFE-free predictions (8273230, 8342852, and 8482121) rank first, showing that these have the most self-similar and regular demand profiles throughout the 92 days among 68 households. The considerable difference between the regularity level of the daily load profiles of households 8273230 and 8487285, whose MAPE and MAPE* values are compared above in Fig. 5, can be seen in Fig. 7, which shows all 92 daily load profiles of those buildings in Fig. 7a and Fig. 7b, respectively. These are complemented by the dendrograms in Fig. 8 that visualise the clustering results of the same houses.

VI. Discussion

Given the strong *within-household* variability of the demand profiles, where as many as 91 different clusters are found for 92 days - it is not surprising that prediction methods fail to provide robust predictions given such variable energy consumption. If this is the case, it should be detected during the prediction evaluation. Nevertheless, the evaluation methods available to the community are not able to do this.

The 1-SS method provides a conceptually straightforward method to detect the PFE in predictions. We note that although the ranking in Table V explicitly reveals the relation between the irregular pattern in data and the PFE, clustering cannot be used as a tool to identify the PFE. This is because, first, hierarchical clustering requires a choice of hyper-parameters such as the threshold value (here: corr=0.75); and second, clustering results rely on dataset features, such as dataset length, granularity and the like. While the ranking of houses is supposed to be independent of these hyper-parameters, the absolute number of clusters is not.

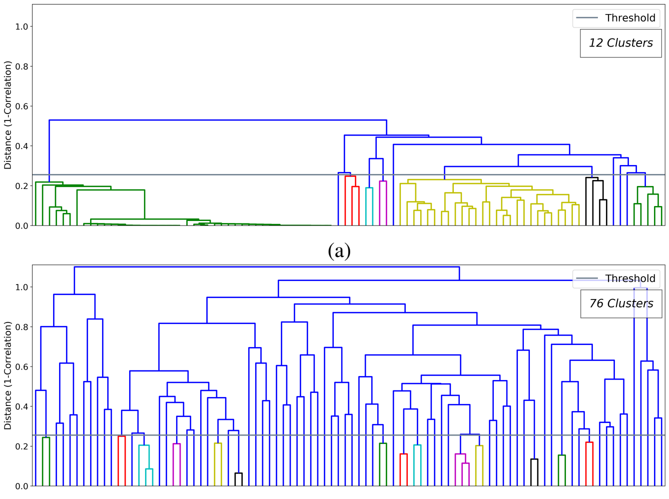
Our empirical regularity analysis focused on the overall variability of demand profiles. A detailed analysis of the underlying reasons why predictions "fall back" on "the most recent observation" is out of the scope of this text. However, the auto-correlation analyses offer to make a progress in this direction. The auto-correlation analysis evaluates the similarity between *observations* from the same time-series variable as a



(a)

(b)

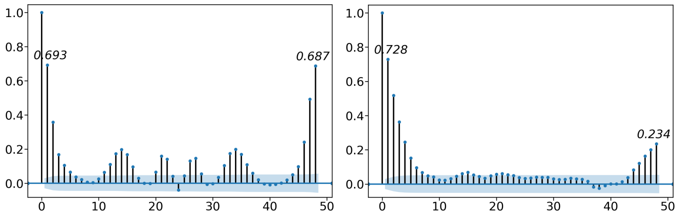
Fig. 7. 92 Daily load profiles of PFE-free House 8273230 (a) and PFE-affected House 8487285 (b).



(a)

(b)

Fig. 8. Dendrograms showing hierarchical clustering result for household 8273230 (a), with PFE-free predictions, and House 8487285 (b), with PFE-affected predictions. Except for blue, each colour bunch represents a cluster and blue lines represent outlier daily profiles.



(a) (b)

Fig. 9. Auto-correlation analyses of PFE-free House 8273230 (a) and PFE-affected House 8487285 (b).

function of the delay between the observations. The distance (delay) between observations is called lag. For example, in a half-hourly dataset, $lag1$ refers to the observation 30 minutes prior to the most recent one, and $lag48$ to the observation 24 hours earlier. Auto-correlation, thus, offers a way to measure and understand similarity *within* a day and sequential days. For instance, Fig. 9a shows significant auto-correlation values at various lags throughout the day for a household whose predictions are PFE-free. In particular, the correlation values at $lag1$ (0.693) and $lag48$ (0.687) are almost equal to each other, illustrating the similarity between the consecutive days. In contrast, Fig. 9b shows the auto-correlation results for a household with PFE-affected predictions. Here, the correlation value at $lag1$ is by far higher than values at all the other lags. In other words, in such a volatile dataset, the current value *is* most correlated to the previous time step. Prediction methods, thus, are consistent to forecast that the future will be "similar to the current".

The objective of our analysis was not to demonstrate that the forecasting methods are "broken" in a previously unknown way. Instead, we aim to guard against overconfidence in prediction outcomes and to mitigate the risk of developing forecast models on datasets that do not work as might be expected. By applying the 1-SS, model developers have a tool to detect the PFE in their predictions. When they do, we recommend a review of the dataset they use and their features. Developers should explore the availability of additional features that are the predicted output is contingent on. In the domain of residential energy demand, these might include the features that determine why, when, and how electrical energy is consumed in a building, including the lifestyle of occupants and daily routines, power consumption of appliances, weather data, and the like. Such augmented datasets hold the potential to include the regularity that forecasting methods rely on and contribute to the reduction of the prediction error.

Nonetheless, historical load data is still an important predictor in the domain of residential load forecasting. In Table VI, we compare accuracy of LSTM RNN for five randomly selected houses with and without historical loads in the input set. Similar to [14], we find that the use of historical data brings substantial improvement in prediction accuracy. Given that historical data is of such high value to prediction accuracy, we propose the 1-SS evaluation method in order to identify when a historical dataset provides insufficient regularity for prediction methods to provide robust output.

We would like to close with a review of the PFE related to the phase error and DPE. The DPE arises from the use of point-wise metrics during the evaluation of discontinuously displaced predictions also known as phase error. The phase error that the DPE is concerned with can occur at any point in time and with varying time steps - backwards or forwards - for all sorts of reasons. The solutions proposed for the DPE work by dropping the time dimension during evaluation, and are applicable to application areas that are tolerant to a displacement of events in time. Not all applications provide this temporal flexibility (such as peak shaving, storage scheduling, and dynamic pricing). As a matter of fact, the PFE refers to predictions systematically trailing the actual loads one step behind in time since they are extrapolated from the most recent load value used in the input set due to irregularity in data. Dropping the time dimension cannot be a solution to PFE. This is because if it was applicable for systematic and continuous prediction displacement, the simplest method, the persistence model introduced in Problem Statement, would always be the most superior method.

TABLE VI
ACCURACY COMPARISON WITH AND WITHOUT HISTORICAL DATA IN INPUT SET.

House ID	Historical Data in Input Set	MAPE	RMSE	Corr
8196671	No	66.988	0.233	0.320
	Yes	28.727	0.122	0.635
8350006	No	35.446	0.204	0.571
	Yes	21.296	0.145	0.751
8432046	No	77.413	0.671	0.466
	Yes	63.266	0.477	0.768
8540084	No	135.099	0.318	0.365
	Yes	78.493	0.226	0.662
9393680	No	172.943	0.527	0.403
	Yes	100.083	0.467	0.557

VII. Conclusion

In this text, we start with the observation that evaluation methods applied to time-series predictions, commonly used for electric load forecasting, fail to detect model-degradation when trained with highly irregular data. We also introduce the PFE, referring to an effect of predictions systematically following the actual load one step behind, which may be detrimental to the final applications that have no tolerance to temporal displacement. We then illustrate the risks associated with this and to mitigate these risks, propose the 1-SS as a PFE detection method. We investigate the use of the 1-SS for single- and multi-step predictions. We finally investigate how irregularity in data causes predictions to be affected by the PFE.

We illustrate the PFE on a real-world dataset of 68 houses by deploying advanced machine learning techniques. We replicate the results of a recent, peer-reviewed study and show that standard evaluation metrics are insufficient to detect the PFE. According to the 1-SS results, only 3 of the houses have PFE-free predictions, whereas predictions of 64 of them are PFE-affected and the remaining household is inconclusive. Finally, through analysis of similarity *between-day* and *within-day* through hierarchical clustering and auto-correlation, we make steps towards a more formal description of the PFE.

As a consequence, the PFE has a strong potential to endanger network security and resilience, as well as the domestic economy. We recommend that model developers apply the 1-SS method to examine the presence of the PFE before deploying models in smart applications. Additionally, in order to overcome the PFE and increase the robustness of residential demand forecasts, we recommend augmenting the input feature set with features that the prediction outputs might be contingent on. As future work, first, we see the evaluation of the PFE in other aspects of power systems and electricity market studies e.g. price forecasting and solar/wind power generation forecasting and second, we see the investigation of whether the PFE can be used for improving the predictions accuracy.

Acknowledgment

The authors gratefully acknowledges that this work was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) - UK Collaboratorium for Research in Infrastructure & Cities: Urban Observatories (Strand B) [EP/P016782/1]. The data used in this study is openly available at "https://data.gov.au/data/dataset/smart-grid-smart-city-customer-trial-data", which is also cited in the References section of this paper.

References

[1] S. Haben, G. Giasemidis, F. Ziel, and S. Arora, "Short term load forecasting and the effect of temperature at the low voltage level," *Int. J. Forecasting*, vol. 35, pp. 1469–1484, 2019.

[2] I. K. Nti, M. Teimeh, A. F. Adekoya, and O. Nyarko-boateng, "Forecasting Electricity Consumption of Residential Users Based on Lifestyle Data Using Artificial Neural Networks," *ICTACT J. Soft Computing*, vol. 10, no. 03, pp. 2107–2116, 2020.

[3] "Eurostat - statistics explained," [Online], https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Energy_consumption_in_households#Energy_products_used_in_the_residential_sector, Accessed: 02.11.2021.

[4] D. L. Marino, K. Amarasinghe, and M. Manic, "Building Energy Load Forecasting using Deep Neural Networks," in *Proc. 42nd Ann. Conf. IEEE Ind. Elect. Soc.*, 2016, pp. 7046–7051.

[5] M. Voss, "Permutation-Based Residential Short-term Load Forecasting in the Context of Energy Management Optimization Objectives," in *Proc. 11. ACM Int. Conf. Future Energy Sys.*, 2020, pp. 231–236.

[6] L. Fan, J. Li, and X. P. Zhang, "Load prediction methods using machine learning for home energy management systems based on human behavior patterns recognition," *CSEE J. Power Energy Sys.*, vol. 6, no. 3, pp. 563–571, 2020.

[7] M. Rowe, T. Yunusov, S. Haben, C. Singleton, W. Holderbaum, and B. Potter, "A peak reduction scheduling algorithm for storage devices on the low voltage network," *IEEE Trans. Smart Grid*, vol. 5, no. 4, pp. 2115–2124, 2014.

[8] W. Kong, Z. Y. Dong, Y. Jia, D. J. Hill, Y. Xu, and Y. Zhang, "Short-Term Residential Load Forecasting Based on LSTM Recurrent Neural Network," *IEEE Trans. Smart Grid*, vol. 10, no. 1, pp. 841–851, 2019.

[9] M. Alhussein, K. Aurangzeb, and S. I. Haider, "Hybrid CNN-LSTM Model for Short-Term Individual Household Load Forecasting," *IEEE Access*, vol. 8, pp. 180 544–180 557, 2020.

[10] W. Kong, Z. Y. Dong, D. J. Hill, F. Luo, and Y. Xu, "Short-term residential load forecasting based on resident behaviour learning," *IEEE Trans. Power Sys.*, vol. 33, no. 1, pp. 1087–1088, 2018.

[11] A. Estebsari and R. Rajabi, "Single residential load forecasting using deep learning and image encoding techniques," *Electronics*, vol. 9, no. 1, pp. 68–85, 2020.

[12] M. Rowe, T. Yunusov, S. Haben, W. Holderbaum, and B. Potter, "The real-time optimisation of DNO owned storage devices on the LV network for peak reduction," *Energies*, vol. 7, pp. 3537–3560, 2014.

[13] C. Deb, F. Zhang, J. Yang, S. E. Lee, and K. W. Shah, "A review on time series forecasting techniques for building energy consumption," *Renew. Sustain. Energy Rev.*, vol. 74, pp. 902–924, 2017.

[14] X. Liu, Z. Zhang, and Z. Song, "A comparative study of the data-driven day-ahead hourly provincial load forecasting methods: From classical data mining to deep learning," *Renew. Sustain. Energy Rev.*, vol. 119, p. 109632, 2020.

[15] Z. Zheng, H. Chen, and X. Luo, "A Kalman filter-based bottom-up approach for household short-term load forecast," *Appl. Energy*, vol. 250, pp. 882–894, 2019.

[16] S. Haben, J. Ward, D. Vukadinovic Greetham, C. Singleton, and P. Grindrod, "A new error measure for forecasts of household-level, high resolution electrical energy consumption," *Int. J. Forecasting*, vol. 30, no. 2, pp. 246–256, 2014.

[17] L. Vallance, B. Charbonnier, N. Paul, S. Dubost, and P. Blanc, "Towards a standardized procedure to assess solar forecast accuracy: A new ramp and time alignment metric," *Solar Energy*, vol. 150, pp. 408–422, 2017.

[18] D. Bunn and E. D. Farmer, "Comparative models for electrical load forecasting," *John Wiley and Sons*, 1985.

[19] M. Lange, "On the uncertainty of wind power predictions - Analysis of the forecast accuracy and statistical distribution of errors," *J. Solar Energy Eng.*, vol. 127, no. 2, pp. 177–184, 2005.

[20] Y. Fujimoto, Y. Takahashi, and Y. Hayashi, "Alerting to rare large-scale ramp events in wind power generation," *IEEE Trans. Sustain. Energy*, vol. 10, no. 1, pp. 55–65, 2019.

[21] L. L. Takacs, "A two-step scheme for the advection equation with minimized dissipation and dispersion errors," *Monthly Weather Rev.*, vol. 113, no. 6, pp. 1050–1065, 1985.

[22] D. Hou, E. Kalnay, and K. K. Droegemeier, "Objective verification of the SAMEX '98 ensemble forecasts," *Monthly Weather Rev.*, vol. 129, no. 1, pp. 73–91, 2001.

[23] M. Cui, B. M. Hodge, J. Zhang, D. Ke, A. Florita, and Y. Sun, "Solar power ramp events detection using an optimized swinging door algorithm," in *Proc. the ASME Design Eng. Tech. Conf.*, 2015, pp. 1–10.

[24] W. Zhu, L. Zhang, M. Yang, and B. Wang, "Solar power ramp event forewarning with limited historical observations," *IEEE Trans. Ind. App.*, vol. 55, no. 6, pp. 5621–5630, 2019.

[25] M. Abuella and B. Chowdhury, "Forecasting of solar power ramp events: A post-processing approach," *Renew. Energy*, vol. 133, pp. 1380–1392, 2019.

[26] C. Ferreira, J. Gama, L. Matias, A. Botterud, and J. Wang, "A survey on wind power ramp forecasting," *Tech. Rep. ANL/DIS-10-13*, 2010.

[27] T. Ouyang, X. Zha, L. Qin *et al.*, "A survey of wind power ramp forecasting," *Energy Power Eng.*, vol. 05, no. 04, pp. 368–372, 2013.

[28] C. Gallego-Castillo, A. Cuerva-Tejero, and O. Lopez-Garcia, "A review on the recent history of wind power ramp forecasting," *Renew. Sustain. Energy Rev.*, vol. 52, pp. 1148–1157, 2015.

[29] T. Zufferey, A. Lepouze, and G. Hug, "Inadequacy of standard algorithms and metrics for short-term load forecasts in low-voltage grids," *IEEE Milan PowerTech, 2019*, pp. 1–6, 2019.

[30] T. Hong and S. Fan, "Probabilistic electric load forecasting: A tutorial review," *Int. J. Forecasting*, vol. 32, no. 3, pp. 914–938, 2016.

[31] D. W. van der Meer, J. Widén, and J. Munkhammar, "Review on probabilistic forecasting of photovoltaic power production and electricity consumption," *Renew. Sustain. Energy Rev.*, vol. 81, pp. 1484–1512, 2018.

[32] S. Arora and J. W. Taylor, "Forecasting electricity smart meter data using conditional kernel density estimation," *Omega*, vol. 59, pp. 47–59, 2016.

[33] J. Lemos-Vinasco, P. Bacher, and J. K. Møller, "Probabilistic load forecasting considering temporal correlation: Online models for the prediction of households' electrical load," *Appl. Energy*, vol. 303, no. 117594, 2021.

[34] J. Munkhammar, D. van der Meer, and J. Widén, "Very short term load forecasting of residential electricity consumption using the Markov-chain mixture distribution (MCM) model," *Appl. Energy*, vol. 282, no. 116180, 2021.

[35] F. Lin, S. J. Liu, H. C. Chao, and J. S. Pan, "Short-term household load forecasting model based on variational mode decomposition and gated recurrent unit with attention mechanism," *J. Netw. Intelligence*, vol. 6, no. 1, pp. 143–153, 2021.

[36] X. Guo, Y. Gao, Y. Li, D. Zheng, and D. Shan, "Short-term household load forecasting based on Long- and Short-term Time-series network," *Energy Reports*, vol. 7, pp. 58–64, 2021.

[37] S. Wang, X. Deng, H. Chen, Q. Shi, and D. Xu, "A bottom-up short-term residential load forecasting approach based on appliance characteristic analysis and multi-task learning," *Elect. Power Sys. Res.*, vol. 196, 2021.

[38] N. M. Ibrahim, A. I. Megahed, and N. H. Abbasy, "Short-Term Individual Household Load Forecasting Framework Using LSTM Deep Learning Approach," in *5th Int. Symp. Multidisc. Studies Innov. Tech.* IEEE, 2021, pp. 257–262.

[39] H. Shi, M. Xu, and R. Li, "Deep Learning for Household Load Forecasting – A Novel Pooling Deep RNN," *IEEE Trans. Smart Grid*, vol. 9, no. 5, 2018.

[40] K. Gajowniczek and T. Zabkowski, "Electricity forecasting on the individual household level enhanced based on activity patterns," *PLoS ONE*, vol. 12, no. 4, pp. 1–26, 2017.

[41] N. Andriopoulos, A. Magklaras, A. Birbas, A. Papalexopoulos, C. Valouxis, S. Daskalaki, M. Birbas, E. Housos, and G. P. Papaioannou, "Short term electric load forecasting based on data transformation and statistical machine learning," *Appl. Sci.*, vol. 11, no. 1, pp. 1–22, 2021.

[42] L. Jiang, X. Wang, W. Li, L. Wang, X. Yin, and L. Jia, "Hybrid Multitask Multi-Information Fusion Deep Learning for Household Short-Term Load Forecasting," *IEEE Trans. Smart Grid*, vol. 12, no. 6, pp. 5362–5372, 2021.

[43] R. Bonetto and M. Rossi, "Machine Learning Approaches to Energy Consumption Forecasting in Households," *arXiv preprint arXiv:1706.09648*, 2017.

[44] L. Li, C. J. Meinrenken, V. Modi, and P. J. Culligan, "Short-term apartment-level load forecasting using a modified neural network with selected auto-regressive features," *Appl. Energy*, vol. 287, 2021.

[45] M. Imani and H. Ghassemian, "Residential load forecasting using wavelet and collaborative representation transforms," *Appl. Energy*, vol. 253, no. 113505, 2019.

[46] A. Rahman, V. Srikumar, and A. D. Smith, "Predicting electricity consumption for commercial and residential buildings using deep recurrent neural networks," *Appl. Energy*, vol. 212, pp. 372–385, 2018.

[47] J.-S. Chou and D.-S. Tran, "Forecasting energy consumption time series using machine learning techniques based on usage patterns of residential householders," *Energy*, vol. 165, pp. 709–726, 2018.

[48] S. Naji, A. Keivani, S. Shamshirband, U. J. Alengaram, M. Z. Jumaat, Z. Mansor, and M. Lee, "Estimating building energy consumption using extreme learning machine method," *Energy*, vol. 97, pp. 506–516, 2016.

[49] W. Yu, D. An, D. Griffith, Q. Yang, and G. Xu, "On statistical modeling and forecasting of energy usage in smart grid," in *Proc. ACM Conf. Res. in Adapt. & Converg. Sys.*, 2014, pp. 12–17.

[50] F. Zhang, C. Deb, S. E. Lee, J. Yang, and K. W. Shah, "Time series forecasting for building energy consumption using weighted Support Vector Regression with differential evolution optimization technique," *Energy Build.*, vol. 126, pp. 94–103, 2016.

[51] Australian Government (data.gov.au), "Smart-grid smart-city customer trial data," [Online], https://data.gov.au/data/dataset/smart-grid-smart-city-customer-trial-data, Accessed: 03.03.2020.

[52] Z. Liu, X. Sun, S. Wang, M. Pan, Y. Zhang, and Z. Ji, "Midterm Power Load Forecasting Model Based on Kernel Principal Component Analysis and Back Propagation Neural Network with Particle Swarm Optimization," *Big Data*, vol. 7, no. 2, pp. 130–138, 2019.

[53] G. F. Fan, Y. H. Guo, J. M. Zheng, and W. C. Hong, "A generalized regression model based on hybrid empirical mode decomposition and support vector regression with back-propagation neural network for mid-short-term load forecasting," *J. Forecast.*, vol. 39, pp. 737–756, 2020.

[54] B. Nepal, M. Yamaha, A. Yokoe, and T. Yamaji, "Electricity load forecasting using clustering and ARIMA model for energy management in buildings," *Japan Arch. Rev.*, vol. 3, no. 1, pp. 62–76, 2020.

[55] V. Cohen-Addad, V. Kanade, F. Mallmann-Trenn, and C. Mathieu, "Hierarchical clustering: Objective functions and algorithms," *J. ACM*, vol. 66, no. 4, pp. 1–42, 2019.

[56] Y. Zhao and G. Karypis, "Evaluation of hierarchical clustering algorithms for document datasets," in *Proc. Int. Conf. Inf. and Knowl. Manag.*, 2002, pp. 515–524.

Huseyin Burak Akyol received the B.Sc. degree in Computer Engineering from Gazi University, Ankara, Turkiye, in 2013, and the M.Sc. degree in Advanced Computer Science (with Distinction) from the University of Leicester, Leicester, U.K., in 2017. He is currently pursuing the Ph.D. with the Department of Computer Science, the University of Bristol, Bristol, U.K. His main research interests include Machine Learning, Big Data, Artificial Intelligence, and Time-Series Analysis.

Chris Preist received the Ph.D. degree in Computer Science from Imperial College, London, U.K., in 1998. He is currently a Professor in Sustainability and Computer Systems in the Department of Computer Science at the University of Bristol, and University Academic Director of Sustainability. His research focuses on the environmental effects of digital technology – both understanding and mitigating the direct negative impacts and applying technology to reduce impacts elsewhere in the economy. Prior to joining the University of Bristol in 2009, he was Head of Sustainable IT Research at HP Labs, Bristol.

Daniel Schien received the Ph.D. degree in Computer Science from the University of Bristol, Bristol, U.K. in 2015, and the M.Sc. degree in Computer Science from Technical University Berlin, Berlin, Germany. He is currently a Senior Lecturer in Computer Systems in the Department of Computer Science at the University of Bristol. His research aims at improving our understanding and mitigating the environmental impact of energy consumption in the built environment with a focus on information and communication technologies (ICT).